Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning

[1]Mudassir Khan
[1]Research Scholar
Noida International University
mkhanmudassir907@gmail.com
mudassirkhan12@gmail.com

[2]Aadarsh Malviya
[2]Assistant Professor
Noida International University
aadarsh.malviya@niu.edu.in

2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.201

***Abstract*—Sentiment analysis acquired a great area of attention in the microblogging websites and analysis of sentiment is a practice of categorization and identification of opinions that are articulated as speech, text, database sources and tweets to detect if opinion is negative, positive or neutral. The challenge lies in determining sentiment from the tweets due to the unique characteristics of Twitter data. This paper presents an approach for sentiment analysis by adapting a Hadoop framework and deep learning classifier. The Hadoop cluster is used for the distribution of data for extracting the features. Then, the significant features are extracted using the twitter data. The deep learning classifier, namely deep recurrent neural network classifier is used assign a real-valued review to each input twitter data thus, classifying the input data into two classes, such as positive review and negative review. The analysis of the performance is done using metrics like, classification accuracy, sensitivity and specificity. In contrast to classical strategies, the proposed method offered improved classification accuracy of 0.9302, better sensitivity of 0.9404 and high specificity of 0.9157, respectively.**

Keywords—Sentiment analysis, feature extraction, feature classification, Hadoop cluster, deep recurrent neural network.

1. INTRODUCTION

Microblogging websites have emerged as the source with various information. The people discuss the current issues, post their opinions on various topics, and express their opinion regarding the products they use in their daily life in microblogs. The manufacturing companies have also started conducting polls to get a sentimental analysis of their product and study the reactions of the users through the microblogs [1]. The analysis of Sentiment is also known as opinion mining, subjectivity examination, and extraction of appraisals. The tasks involved in sentimental analysis are extraction of sentiment, classification of subjectivity, classification of sentiment, summarization of opinions and detection of spams. The Sentiment analysis analyzed the emotion, attitude and sentiment of the people towards the elements, like individuals, products, organizations, topics, and services [2]. The development of big data has provided enormous opportunities in the field of sentiment analysis [3]. Sentiment analysis provided information regarding the product to the customers before buying it. This analysis data is used by the firms and Marketers to understand about their services or products that help in offering services as per the requirements of the user.

In the past days, analysis of sentiment using twitter data has gained substantial interest. Unified methods are developed for analyzing the tweets with importance on a domain□specific model [4]. The sentiments are categorized using machine learning techniques or supervised approach and lexicon based strategies or the unsupervised approach. The machine learning approach uses Support Vector Machine (SVM) [5], k-nearest neighbors (k-NN) [6], and Naïve Bayes (NB) [7] for sentiment classification. Sentiments are also anticipated using expressions, words, documents, patterns or Natural Language Processing (NLP) [8]. Several studies have been conducted for analyzing the user sentiments in Twitter. The user reviews in twitter are classified as positive, negative or neutral and they are categorized into two different classes, such as white box and black box. The black box classification algorithm uses definite knowledge-based rules for addressing the issues of classification. Some of the black box classification algorithm are SVM, NB, artificial neural networks (ANNs), maximum entropy (ME), discriminant analysis (DA), k-NN and genetic algorithms (GAs) [9, 10]. The strategies based on white-box classification includes set-based strategies [11], decision trees, associative rules and sequential covering [8] algorithms.

The objective of the proposed sentiment analysis model is to allocate a review of real-valued input twitter data using deep recurrent neural network. The review specifies the expediency of result for analyzing sentiments. The proposed approach involves two steps, namely extraction of features, and classification. Initially, the input twitter data is subjected to the Hadoop cluster to distribute data for the extraction of feature. The obtained features are fed to classification module, wherein a classifier, named deep recurrent neural network is employed. On the basis of extracted features, the deep recurrent neural network classifier performs the classification, providing two classes, namely positive review and negative review

The paper is structured as: section 1 illustrates sentiment analysis; Section 2 portrays the survey of classical sentiment analysis strategies. Section 3 elaborates proposed method, section 4 instantiates results and Section 5 provides the conclusion.

II. LITERATURE SURVEY

The literature survey of sentiment analysis strategies are as follows: Asghar, M.Z.,*et al*.[4] developed a Twitter sentiment analysis framework known as T□SAF. This method focused on domain□specific words from various domains thus, improving performance of the sentiment classifier. However, this method failed to include context□aware features for tweets classification. Asghar, M.Z.*et al*.[9] modeled a Rule Induction Framework known as RIFT. Although this method improved the performance by classifying the tweets in accordance with emoticons and slang, it resulted in inaccurate scoring.

Plunz, R.A.*et al*.[12] designed a Geolocated Twitter framework that helped the people to understand the expressed sentiment. However, this method failed to deal with densely populated metropolitan cities. Rodrigues, A.P. and Chiplunkar, N.N [13] developed a Hybrid Lexicon-Naive Bayesian Classifier (HL-NBC) for analyzing the sentiments of twitter data. Although this method had low execution time, it failed to filter the sarcasm sentiments and classify them.

A. Challenges

- The sentiment analysis based on Twitter analyzed the positive, neutral and negative reviews. The major challenge lies in building technology that identified and compiled the overall sentiment [14].
- The creation of noise while labelling the data is one of the challenges faced during the sentiment analysis of Twitter data [13].
- In [4], abbreviations and slang, limited lexicons of emoticons, insufficient and irregular words expressed by the users in their post resulted in low classifier accuracy in detecting the polarity of the tweets and the incomplete coverage of domain□specific words had resulted in incorrect sentiment classification and scoring.
- Inorder to cleanse and analyze the sentiments of user at satisfactory level only a limited automatic and sophisticated Twitter-based content analysis tools were accessible. Hence, a dedicated and integrated platform based on Twitter-based content was needed for extracting the obtainable information in public from a huge text streams to synthesize and analyze the feedback of the customer [9].

III. Proposed Classification Framework Based on Hadoop for Sentiment Analysis

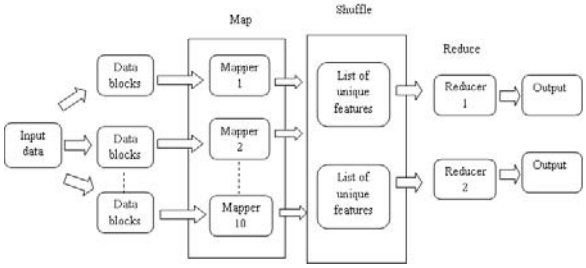


Figure 1. portrays the block diagram of analyzing sentiment

Sentiment analysis is a procedure for analyzing and classifying the opinion from text data. The sentimental analysis provided an overview of the public opinion about the certain topics. Here, the sentiment analysis is done using the Hadoop framework and deep learning classifier. Initially, the input twitter data is subjected to Hadoop cluster to distribute data for the extraction of features. The extraction of feature is carried out in the mapper phase. In the mapper phase, the significant features, like all-caps, emoticon, hashtag, elongated units, sentiment lexicon, negation, and punctuation are extracted from the twitter data. The obtained features are then fed in the shuffle, where lists of unique features are selected. The list of unique features is fed to the reducer. In the reducer phase, the features are classified using deep recurrent neural network classifier, which classifies the features into two classes, namely positive review and negative review [15]. Figure 1 portrays the schematic view for analyzing sentiments.

A.Hadoop framework for sentiment analysis

The Hadoop framework handles the issues with parallel executions, like recovery between the tasks, detection of failure and automatic tasks synchronization and dealt with large-scale data processing. In Hadoop, the data are processed by the user using MapReduce models. Thus, the Hadoop framework is used for effective analysis of sentiment using twitter data.

a)Mapper phase in Hadoop framework

In the initialization phase of the Hadoop framework, 100 reviews from the twitter are considered as the input data. The number of Hadoop Distributed File System (HDFS) blocks required to process the input data are determined and HDFS path to the block are computed. The computed block paths are then copied using configuration file and job jar file to a shared directory of HDFS. The Map tasks required to process all the blocks are created using framework. Here, 10 such Mappers are required for processing the data blocks. The data blocks are processed and the features, like all-caps, emoticon, hashtag, elongated units, sentiment lexicon, negation, and punctuation are extracted from the twitter data. The all-caps represent count of words considering all characters in the upper case. The emoticon features indicates the presence of the positive or negative emoticon. The presence of the number of hash tag is indicated by the Hashtag feature. The elongated unit indicates the number of basic computational that contains the elongated words. Sentiment lexicon is the count of sentiment words, total sentiment score, score of last sentiment words, and the maximal sentiment score for each lexicon. Negation indicates the number of negations as individual units in the segmentation. Punctuation is the number of contiguous sequence of question mark, dot and exclamation mark [16].

b) Shuffle phase in Hadoop framework

The intermediate outcome of Map phase was passed to the shuffle phase prior to the reduce phase. The shuffle phase has three phases, post-Map shuffling, copying and sorting phase. The post-Map shuffling stage happens at the end of the Map task in which the output of the Map task are divided and arranged. The features extracted from the twitter data are partitioned and sorted in this stage. In the copying stage, the Map tasks outputs is copied to local disk. The sorting stage forms the input to the reducer phase by performing merging and sorting operation. During the merging operation, the features that are unique are merged and sorted, which is given as the input to the reducer phase.

c) Reduce phase in Hadoop framework

In the reducer phase, the features are classified as positive and negative review using deep RNN. RNN is a supervised machine learning approach that is made up of artificial neurons having two or more feedback loops.

RNN requires targeted input pairs and training database for training in supervised manner. RNN comprises input layer, recurrent hidden layer and the output layer. The sequence of vectors at time t is given as the input to the network,

$$H_t = (H_1, ..., H_n) \quad (1)$$

The input is given to the hidden units of the fully connected RNN for which the connections are represented as, m_{uv}, a weight matrix. The recurrent connections are used for joining the h hidden units of the hidden layer.

$$W_t = (D_1, D_2, ..., D_m) \quad (2)$$

The hidden layer indicates the state space of the system, which is represented as,

$$W_t = M_\alpha(B_t) \quad (3)$$

where, the activation function of the hidden layer is denoted as, $M_\alpha(.)$ and the output gate (B_t) determined the exposed memory content, which is expressed as,

$$B_t = \gamma_{uv} H_t + \gamma_{vv} D_{t-1} + b_{hidden} \quad (4)$$

where, γ is the weight matrix and b_{hidden} is the bias of hidden units. The hidden layers linked with the hidden units are represented as, γ_{vv}. The output layer of the deep RNN is given as,

$$V_t = V_1, ..., V_n \quad (5)$$

where, $n = 2$, At time t, the output layer is represented as,

$$V_t = a_o(\gamma_{kh} D_t + b_{output}) \quad (6)$$

where, $a_o(.)$ is the activation function and b_{output} is the bias vector of the input layer. where, γ_{vl} is the weighted connection of the hidden units that are linked to the output layer. The above equation is repeated for time, $t = (1, ..., \tau)$. The output of the output layer is predicted by the hidden layer using the input vector in every time step. The unique information of the past states is summarized as the set of values in the hidden state RNN. The accurate classification in the output layer is provided by the combined information.

IV. Results and Discussion

This section explains the analysis of proposed Hadoop based deep RNN method. The effectiveness of the proposed Hadoop based deep RNN method is determined by comparative analysis of the proposed Hadoop based deep RNN using conventional strategies.

A. Experimental setup

The proposedHadoop based deep RNN is executed in PYTHON software that executes in PC with Windows 8 OS. Twitter US Airline Sentiment dataset [17] is used for the analysis of the proposed Hadoop based deep RNN method.

B. Performance metrics

The performance metrics, like classification accuracy, sensitivity and specificity are considered for the analysis of the proposed Hadoop based deep RNN method.

C. Comparative methods

The proposed Hadoop based deep RNN is evaluated with the classical strategies, like Rough Set Theory [18], PSO [19] and NB [20].

D. Comparative analysis

The analysis of the proposed Hadoop based deep RNN method is performed considering performance measures, like classification accuracy, sensitivity and specificity**.**

a)Analysis using classification accuracy: Figure 2 depicts the analysis of the classification accuracy by varying the training percentage**.** Figure 2 a) portrays analysis of classification accuracy for Number of Mappers=5. For the training percentage 90, the classification accuracy obtained by the Rough Set Theory, PSO, NB and the proposed Hadoop based deep RNN method is 0.9086, 0.9118, 0.9206, and 0.9300 respectively. Figure 2 b) shows the analysis of theclassification accuracy for Number of Mappers=6. When the training percentage is 90, classification accuracy obtained by the Rough Set Theory, PSO, NB and the proposed Hadoop based deep RNN method is 0.9063, 0.9100, 0.9232 and 0.9302 respectively.

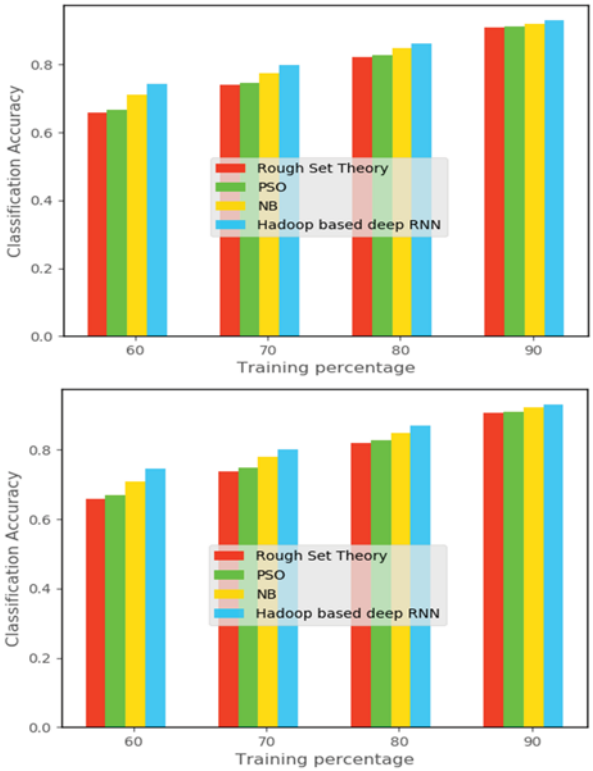


Figure 2. Analysis using classification accuracy a) Number of Mappers=5 b) Number of Mappers=6

b)Analysis based on Sensitivity: Figure 3 depicts the analysis of the sensitivity by varying the training percentage**.** Figure 3 a) shows the analysis of thesensitivity for Number of Mappers=5. When the training percentage is 90, sensitivity obtained by the Rough Set Theory, PSO, NB and the proposed Hadoop based deep RNN method is 0.9214, 0.9242, 0.9320 and 0.9403 respectively. Figure 3 b) portrays analysis of sensitivity for Number of Mappers =6. For the training percentage 90, the sensitivity obtained by the Rough Set Theory, PSO, NB and the proposed Hadoop based deep

RNN method is 0.9194, 0.9227, 0.9344 and 0.9404 respectively.

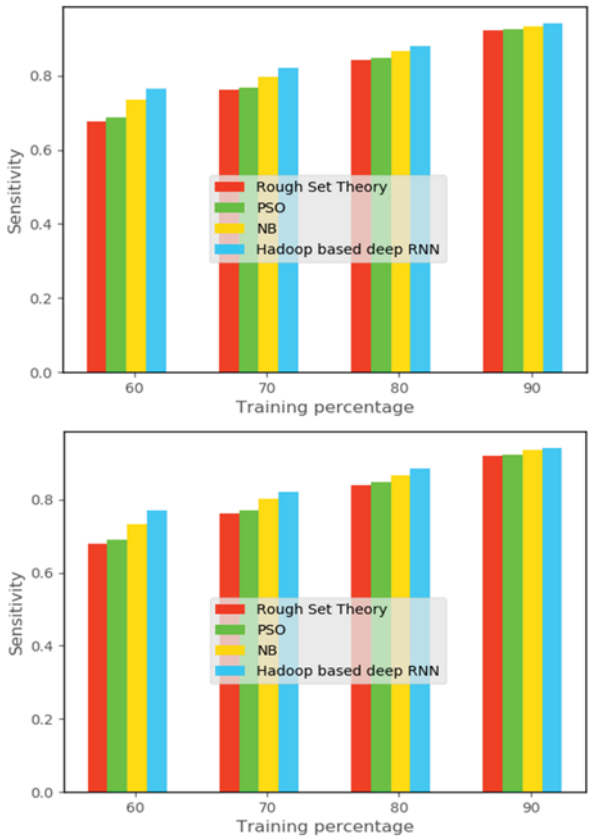
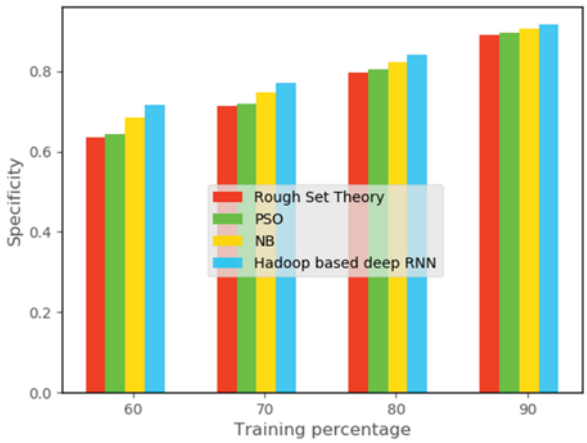


Figure 3. Analysis using sensitivity a) Number of Mappers =5 b) Number of Mappers =6

c)Analysis using Specificity: Figure 4 portrays analysis of specificity by varying the training percentage**.** Figure 4 a) illustrates analysis of the specificity for Number of Mappers=5. For the training percentage 90, the specificity obtained by the Rough Set Theory, PSO, NB and the proposed Hadoop based deep RNN method is 0.8908, 0.8944, 0.9045 and 0.9155 respectively. Figure 4 b) shows the analysis of thespecificity for Number of Mappers =6. When the training percentage is 90, specificity obtained by the Rough Set Theory, PSO, NB and the proposed Hadoop based deep RNN method is 0.8881, 0.8924, 0.9076 and 0.9157 respectively.



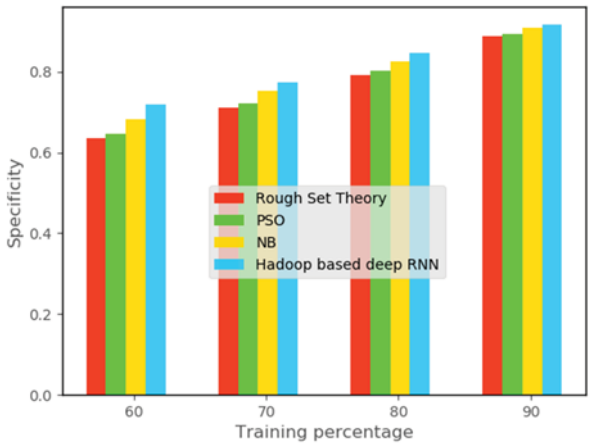


Figure 4. Analysis using specificity a) Number of Mappers =5, b) Number of Mappers =6

V. Conclusion

The analysis of sentiments using twitter data has gained substantial attention from the online clients, who desired to obtain quick response about the companies or products. This work presents a Hadoop framework and deep learning classifier for sentiment analysis. The Hadoop framework distributes data for feature extraction process. In feature extraction process, the significant features, like all-caps, emoticon, hashtag, elongated units, sentiment lexicon, negation, and punctuation is extracted using input twitter data. On the basis of extracted features, the deep learning classifier, namely the deep recurrent neural network is employed in the classification module for classifying input twitter data into two classes namely, positive review and negative review. The analysis is done for the training percentage and the performance of the proposed Hadoop based deep RNN method is evaluated using the performance metrics, such as classification accuracy, sensitivity and specificity. When compared to other existing methods, the proposed Hadoop based deep RNN method provided maximal accuracy, sensitivity and specificity of0.9302, 0.9404 and 0.9157 respectively. The future enhancement can be done by including more features in the feature extraction process.

References

[1]. Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., "Sentiment analysis of twitter data," In Proceedings of the Workshop on Language in Social Media, pp. 30-38, 2011.

[2]. Kharde, V. and Sonawane, P., "Sentiment analysis of twitter data: a survey of techniques," arXiv preprint arXiv:1601.06971, 2016.

[3]. El Alaoui, I., Gahi, Y. and Messoussi, R., "Big Data Quality Metrics for Sentiment Analysis Approaches," In Proceedings of the 2019 International Conference on Big Data Engineering, ACM, pp. 36-43, 2019.

[4]. Asghar, M.Z., Kundi, F.M., Ahmad, S., Khan, A. and Khan, F., "T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme," Expert Systems, vol.35, no.1, pp.12233, 2018

[5]. B. Pang, L. Lee, S. Vaithyanathan, and S. Jose, “Thumbs up?:sentimentclassification using machine learning techniques,” in Proceedings of theACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 79–86, 2002.

[6]. R. Alfred, W. W. Yee, Y. Lim, and J. H. Obit, “Factors affectingsentiment prediction of malay news headlines using machine learningapproaches,” Commun. Comput. Inf. Sci., vol. 652, pp. 289–299, 2016.

[7]. J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive andnegative links in online social networks,” in Proceedings of the 19thinternational conference on World wide web, pp. 641–650, 2010.

[8]. Cohen, W.W, “Fast effective rule induction,” In Proceedings of theTwelfth International Conference on Machine Learning, pp. 115– 123, 1995.

[9]. Asghar, M.Z., Khan, A., Khan, F. and Kundi, F.M., "Rift: a rule induction framework for twitter sentiment analysis," Arabian Journal for Science and Engineering, vol.43, no.2, pp.857-877, 2018

[10]. A. Lin, “Improved Twitter Sentiment Analysis Using Naive Bayes andCustom Language Model,” arXivPrepr. arXiv1711.11081, 2017.

[11]. Chan, C.C. and Liszka, K.J., "Application of rough set theory to sentiment analysis of microblog data," In Rough Sets and Intelligent Systems-Professor Zdzisław Pawlak in Memoriam, Springer, pp. 185-202, 2013.

[12]. Plunz, R.A., Zhou, Y., Vintimilla, M.I.C., Mckeown, K., Yu, T., Uguccioni, L. and Sutto, M.P., "Twitter sentiment in New York City parks as measure of well-being," Landscape and Urban Planning, vol.189, pp.235-246, 2019.

[13]. Rodrigues, A.P. and Chiplunkar, N.N., "A new big data approach for topic classification and sentiment analysis of Twitter data," Evolutionary Intelligence, pp.1-11, 2019.

[14]. Malik, M., Naaz, S. and Ansari, I.R., "Sentiment Analysis of Twitter Data Using Big Data Tools and Hadoop Ecosystem," In proceedings of International Conference on ISMAC in Computational Vision and Bio-Engineering, Springer, pp. 857-863, 2018.

[15]. Tan, Y.S., Tan, J., Chng, E.S., Lee, B.S., Li, J., Date, S., Chak, H.P., Xiao, X. and Narishige, A., 2013. Hadoop framework: impact of data organization on performance. Software: Practice and Experience, 43(11), pp.1241-1260.

[16]. Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, & Ming Zhou, “ A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.11, pp.1750–1761, 2015.

[17]. Twitter US Airline Sentiment, "https://www.kaggle.com/crowdflower/twitter-airline-sentiment", Accessed on September 2019.

[18]. Asghar, M.Z., Khan, A., Khan, F. and Kundi, F.M., "Rift: a rule induction framework for twitter sentiment analysis," Arabian Journal for Science and Engineering, vol.43, no.2, pp.857-877, 2018.

[19]. Nagarajan, S.M. and Gandhi, U.D., "Classifying streaming of Twitter data based on sentiment analysis using hybridization," Neural Computing and Applications, vol.31, no.5, pp.1425-1433, 2019

[20]. Rodrigues, A.P. and Chiplunkar, N.N., "A new big data approach for topic classification and sentiment analysis of Twitter data," Evolutionary Intelligence, pp.1-11, 2019.

[21]. Khan, M. and D. B. Kalra (2018). "AN INSPECTION ON BIG DATA COMPUTING." International Journal of Engineering & Science Research(Special Issue/Article No-52): 326-329.